PRESENTATION DU LOGICIEL SANDCO "SYSTEME POUR L'ANALYSE DES DONNEES COLLECTIVES"

Pierre CHEVAUGEON
Patrice VIMARD

Introduction

Cette communication se propose de présenter les principaux aspects d'un logiciel "pour l'analyse des données collectives" SANDCO, dont l'objectif essentiel est de constituer des fichiers "unité collective" à partir des fichiers "individu" communément disponibles[1].

En effet, le manque de données sur les unités collectives (ménage, famille, noyau familial...) directement utilisables est, depuis longtemps, l'un des principaux obstacles au développement des recherches en "démographie de la famille" dans les pays en développement. Cet obstacle est d'autant plus regrettable que l'information de base, constituée par "les relations de parenté des individus à l'intérieur de leur structure collective d'appartenance", existe le plus souvent dans les fichiers informatiques constitués après chaque enquête. Mais en l'état des outils disponibles, l'utilisation de ces informations pour la création de fichier sur les unités collectives nécessite soit l'élaboration d'un programme informatique spécifique (et par conséquent la connaissance solide d'un langage de programmation ou le recours à un informaticien), soit la mise en oeuvre de certains ordres des logiciels existants, souvent très contraignants à utiliser.

---

[1] La rédaction de ce logiciel est réalisée dans le cadre d'un programme sur l'évolution des structures familiales en Afrique, associant actuellement l'ORSTOM, le CEPED et des institutions africaines : Direction de la Statistique de Côte d'Ivoire, l'ENSEA d'Abidjan et l'URD du Togo, avec le soutien financier du ministère français de la Coopération.

Le logiciel SANDCO a été créé pour palier à ces contraintes. Il comprend un module principal qui permet de constituer un enregistrement pour chaque unité collective, à partir des informations recueillies dans les enregistrements des individus composant l'unité ; il se compose également de différents modules utilitaires de contrôle, d'appariement des fichiers et d'édition des dossiers et d'un module parallèle de recodification ; modules que nous allons présenter brièvement après avoir signalé quelques spécifications techniques de ce logiciel[2].

I. Spécifications techniques et principes d'utilisation

Le logiciel fonctionne sur micro ordinateur de type PC ou compatible comportant une mémoire de 512 K octets au minimum, un disque dur, le système d'exploitation MS/DOS version 2.1 ou suivantes. Il requiert au moins 500 K octets disponibles en plus des fichiers de données.

Il a été développé à l'aide du gestionnaire de bases de données DBASE III plus, du compilateur CLIPPER et du langage de programmation QUICK BASIC.

Ce logiciel peut traiter deux types de fichiers de données : des fichiers au format DBASE III et des fichiers au format "texte ASCII" au sens de MS/DOS. Dans le cas particulier de ce logiciel, les fichiers de données doivent être constitués d'enregistrements de longueur fixe, correspondant chacun à une ligne de questionnaire (ou à un questionnaire) ; tous les enregistrements auront donc la même structure en ce qui concerne la position et la longueur de chacun des champs[3].

---

2 Cette communication n'a pas pour objectif d'être un résumé du manuel provisoire d'apprentissage et d'utilisation (Chevaugeon et Vimard, 1990) ou un exposé didactique de son maniement, mais simplement de présenter les objectifs et les possibilités du système.

3 Pour la plupart des traitements, le logiciel considère que les fichiers sont triés dans l'ordre croissant des identifiants, qu'il s'agisse des identifiants "unité collective" et des identifiants individuels.

L'utilisation des différentes possibilités du logiciel se fait par choix successifs dans une série de menus ; options qui s'établissent en frappant le numéro de la fonction que l'on souhaite voir se dérouler et qui s'éxécute immédiatement (voir comme exemple en annexe le menu principal). Ces choix permettent de se déplacer dans une arborescence afin de réaliser les différentes étapes du traitement souhaité[4].

La logique générale de fonctionnement du système repose sur la notion d'application. En conséquence, la séquence logique de chaque traitement comprend les phases successives suivantes :

a) **gestion des applications** qui permet de définir[5] une application (par son nom, la date de sa création, le nom du fichier de données auquel elle doit être appliquée ;

b) **choix d'une application** qui permet de distinguer l'application qui va être utilisée pour le traitement (parmi toutes celles ayant été préalablement définies);

c) **gestion des tables** où l'on paramètre les différentes règles de contrôle, de constitution des fichiers "unité collective" et d'appariement qui seront utilisées dans les étapes suivantes ;

d) **contrôle des données**, qui permet la recherche de différents types d'erreur dans le fichier de données ;

---

[4] A tout moment la première ligne du haut de l'écran indique la position de l'utilisateur dans l'arborescence ; par exemple la ligne V.1 signifie que l'utilisateur a choisi la fonction V du menu principal (Analyse et appariement) et la fonction 1 du menu V (constitution d'un fichier ménage).

[5] Le processus de constitution d'une nouvelle application comporte la création de 5 fichiers au format DBASE III qui comprendront les différentes règles de contrôle des données lorsque celles-ci auront été définies dans la phase correspondante.

e) **analyse**, nom donné à la phase de création des fichiers "unités collectives"[6] ;

f) **appariement**, qui permet d'associer les enregistrements relatifs à la même unité statistique situés dans deux fichiers[7] ;

g) **édition des dossiers** : étape d'édition de l'ensemble des éléments d'une application : contenus des règles des différentes tables utilisées, caractéristiques des fichiers...

II. Le module d'analyse ou de création des fichiers "unité-collective"

Le module dit d'analyse, qui permet la création des fichiers "unité collective", est la partie centrale du système.

Il autorise la constitution d'un enregistrement pour chaque unité collective dont l'existence a été décelée dans le fichier individuel en passant en revue la séquence des identifiants[8] ; la série des enregistrements de chacune des unités collectives constituant le fichier "unité collective".

Avant de considérer les différents éléments que peut contenir les enregistrements "unité collective", on doit indiquer que la variable stratégique pour la constitution de ce fichier est la variable **"lien de parenté avec le chef de l'unité**

---

[6] Pour une raison de nombre de caractères des lignes des menus les "fichiers unité collective" ont été dénommés dans ces menus "fichier ménage".

[7] Les différentes phases de contrôle, d'analyse et d'appariement, que l'on peut qualifier d'opérationnelles, peuvent être utilisées indépendamment les unes des autres et seules les trois premières étapes (gestion et choix d'une application, gestion des tables) sont obligatoires avant de passer à ces phases opérationnelles.

[8] Cette recherche est faite en partant du principe que chaque enregistrement individuel est caractérisé par le numéro d'identification de son unité collective d'appartenance et par son propre numéro d'ordre dans cette unité.

collective". Cette variable est en effet tout naturellement utilisée pour distinguer le chef de l'unité ainsi que les différentes catégories de parenté qui permettront de définir le type de l'unité et de ses composants[9].

En utilisant cette "variable-clé" l'analyse crée un enregistrement pour chaque unité collective, qui comprend, outre l'identifiant de l'unité, différentes variables sélectionnées ou construites au gré de l'utilisateur. Ces variables peuvent être groupées selon trois catégories :

a) **les variables relatives à un individu déterminé** (le chef de l'unité le plus souvent) ; cette possibilité permet de disposer dans l'enregistrement "unité collective" de tout ou partie des variables caractérisant le chef de cette unité.

b) **les variables relatives à chaque individu** membre de l'unité collective ; ceci permet de disposer dans l'enregistrement "unité collective" de certaines variables des individus composant l'unité (par exemple le niveau de scolarisation, l'activité...).

c) **les variables caractéristiques de la composition et de la structure de l'unité collective** : taille totale, effectif de chacune des différentes catégories de parenté qui ont été définies, type de l'unité, type de deux composants. Ce dernier groupe de variable est le plus intéressant pour l'analyse des structures familiales et le principe de leur constitution mérite d'être précisé.

La variable "taille totale" est sans équivoque ; indiquons seulement que deux variables filtres, la parenté et la résidence, permettent de prendre seulement en compte dans ce calcul certaines catégories d'individus[10]. Le contenu des autres

---

9 Une seconde variable est également importante : la situation de résidence qui fournit un filtre utilisable afin d'effectuer une analyse sur une sous-population du "fichier individu" (les résidents de droit ou de fait par exemple).
10 L'emploi de ces variables filtres détermine également les

variables est directement déterminé par les principes de recodification de la parenté que l'utilisateur définit pour chaque application. En effet, la variable parenté est généralement codifiée sur un caractère (avec 10 modalités au maximum) ou sur deux caractères (avec 100 modalités au maximum) et il peut être utile ou même indispensable[11] d'effectuer des regroupements, selon de grandes catégories de parenté, des liens enregistrés lors de la collecte et de la codification. Cette recodification est l'un des paramètres essentiels de l'analyse ; elle détermine les parentés pour lesquelles seront calculés les effectifs, dans chaque unité, et fournit les fondements de la constitution des typologies caractéristiques des unités collectives.

En effet, pour chacune des parentés définies dans l'analyse, une variable "effectif des individus de cette parenté" est calculée, ce qui permet d'avoir par exemple dans chaque enregistrement "unité collective", le nombre de conjoints, d'enfants, de petits-enfants, de collatéraux, etc.

Pour ce qui concerne les typologies caractéristiques de l'unité collective, trois variables peuvent être créées par le logiciel: type de l'unité (dans son ensemble), type du premier composant, type du second composant. Il suffit pour cela à l'utilisateur de définir les parentés qui entrent dans la composition de l'unité et de ses deux composants. Les variables "type" seront alors calculées, indiquant, à l'aide d'un code à progression géométrique, les parentés effectivement présentes.

Prenons un exemple pour illustrer notre propos. L'utilisateur a choisi 8 catégories de parenté, exprimées par rapport au chef, pour une unité collective définie comme un "ménage" (chef, conjoints, enfants, ascendants, petits-enfants,

---

individus qui sont pris en compte dans la construction des variables "effectif selon la parenté", "types d'unité" et "type de composant".

[11] Remarquons notamment que l'emploi d'une parenté avec 100 modalités conduit à une typologie des unités collectives où le nombre de modalités possibles égale 2 , ce qui est inutilisable; avec 10 modalités le nombre de modalités de la typologie atteint déjà 512 (en considérant qu'il y a au moins un chef par unité).

collatéraux, parents par alliance, sans parenté); il a également choisi 3 catégories pour le premier composant (chef, conjoints, enfants), qualifié de "nucléaire", et 4 catégories pour le second composant qualifié "d'apparenté" (ascendants, petits-enfants, collatéraux, parents par alliance). Les variables "effectif" de chaque catégorie de parenté, "type de ménage", "type du composant nucléaire", "type du composant apparenté" seront alors créées et calculées pour chaque unité collective. Comme illustration, signalons que la variable "type du composant nucléaire" aura 4 modalités qui seront :

1. Chef seul
3. Chef + conjoint(s)
5. Chef + enfant(s)
7. Chef + conjoint(s) + enfant(s)

III. Les modules complémentaires

Nous présenterons dans ce chapitre les modules complémentaires que ceux-ci soient destinés à être utilisés en amont du module central (recodification, contrôle des données) ou en aval de celui-ci (appariement de fichiers, édition de dossiers)[12]

III.1. Recodification des données

Pour une raison de place en mémoire centrale, le module de recodification des données est un module indépendant situé parallèlement au système. Il permet de recoder des variables à un ou deux caractères numériques.

Ce module donne ainsi la possibilité à l'utilisateur d'effectuer une recodification des données individuelles nécessaires au traitement du programme d'analyse proprement dit, tout en conservant intact le "fichier individu" d'origine.

---

[12] Quel que soit leur enchaînement le plus logique, remarquons que ces modules peuvent être utilisés indépendamment les uns des autres; ainsi notamment les modules de contrôle et d'appariement peuvent être employés pour d'autres destinations qu'une exploitation de "données familiales".

III.2. Le module de contrôle

Ce module permet des contrôles, d'une part, des données des enregistrements individuels et, d'autre part, de la séquence des enregistrements des individus de chaque unité collective.

Le contrôle des données individuelles autorise celui des champs de longueur 1, des champs de longueur 2 (recherche de valeur qui n'existe pas dans la table de chiffrement) et de la cohérence entre deux champs (recherche d'incompatibilité entre les valeurs des deux variables d'un même enregistrement)[13].

Le contrôle inter-enregistrements permet de vérifier deux règles distinctes. La première est celle "d'unicité des valeurs"; elle vérifie que, dans une même unité collective, un individu, et un seul, possède une valeur donnée pour la variable choisie. Cette règle permet de contrôler par exemple qu'il existe pour toute unité un chef et un seul.

La seconde règle détecte les doubles et les manquants dans une séquence d'identifiant. Elle permet de vérifier que, pour une même unité collective (définie par son identifiant), la séquence de numérotation des individus (distingués par les numéros d'ordre individuels) est respectée c'est-à-dire qu'il n'existe pas d'enregistrements individuels en double ou manquants.

III.3. Le module d'appariement des fichiers

Ce module consiste à rapprocher 2 fichiers pour en produire un troisième[14]. Chaque enregistrement du fichier résultat peut être constitué de tout ou partie de chacun des enregistrements des fichiers qui sont appariés.

Différents types d'appariement sont d'autre part possibles selon la nature des "fichiers-origine" et selon le type du "fichier-résultat" souhaité par l'utilisateur. Un système d'option permet en effet de définir tout d'abord la nature des

13 Les règles de cohérence entre 2 champs sont énoncées sous la forme "si un champ a une certaine valeur, alors un autre champ ne peut avoir que les valeurs indiquées comme correctes".

14 Par itération il est évidemment possible de rapprocher n fichiers en un seul.

deux fichiers à apparier; chacun d'eux pouvant être un fichier de données individuelles ou un fichier de données agrégées (un fichier "ménage" par exemple). Il permet ensuite de définir le contenu du fichier résultat qui peut être la somme des deux fichiers initiaux, leur intersection ou toute autre combinaison possible.

Par exemple dans le cas d'une enquête à deux passages, l'appariement des fichiers individuels, constitués pour chacun de ces passages, peut créer par exemple des fichiers comprenant:

a) tout individu présent[15] dans au moins l'un des fichiers (fichier-somme);
b) les seuls individus présents à la fois dans les deux fichiers (fichier-intersection);
c) les individus présents aux deux passages à la fois et les individus présents au premier passage (suivi de la population initiale sans considérer les entrées dans l'échantillon entre les passages).

Toute autre combinaison peut être réalisée et l'appariement peut également concerner les deux fichiers-ménage constitués à partir de chaque fichier-individu, le fichier-ménage et le fichier-individu relatifs à un passage...

III.4. Le module d'édition des dossiers

Ce module a été conçu afin de fournir à l'utilisateur un dossier complet dans lequel on trouve une description du fichier initial de données[16], l'ensemble des contrôles réalisés, les différentes analyses et appariements qui ont été effectués avec la description des fichiers obtenus.

---

15 Le terme "présent" ne fait pas ici référence à une situation de résidence mais seulement à la présence physique d'un enregistrement dans un fichier de données.

16 Si le fichier concerné par l'application est un fichier au format DBASE III, c'est la description de ce fichier qui se trouve automatiquement reprise sinon c'est à l'utilisateur d'effectuer la description de son fichier initial de données comme pour tout autre fichier de format DBASE.

Conclusion

Le logiciel SANDCO a été écrit dans le but de constituer un système relativement complet permettant de résoudre les différents problèmes informatiques relatifs au traitement de données sur les unités collectives. Une première version du logiciel et du manuel d'utilisation est actuellement testée par trois équipes de démographes. Lorsque ce test sera achevé et que les modifications jugées nécessaires auront été apportées, le logiciel pourra être largement diffusé. Des développements ultérieurs sont envisagés, notamment sur le traitement des sous-unités, par exemple les différents noyaux familiaux d'un ménage.

Soulignons enfin qu'une bonne connaissance du logiciel permet de l'utiliser pour d'autres applications que le traitement des données collectives. Il est en effet susceptible de "résumer" et de dégager une typologie, sur un enregistrement unique, des informations situées sur une suite d'enregistrements de même nature et relatifs à une même "unité statistique source", par exemple les naissances ou des grossesses d'une femme, les migrations ou les activités d'un individu, les résidences d'un ménage, les parcelles d'une exploitation agricole...

REFERENCES

Chevaugeon P., Vimard P. (1990), "Sandco, Système pour l'Analyse des Données Collectives. Logiciel et manuel d'utilisation", version provisoire, CEPED-ORSTOM, Paris, octobre 1990.

ANNEXE: MENU PRINCIPAL

```
================================================================
             CONTROLES ET ANALYSES DES FICHIERS
                        MENU PRINCIPAL
================================================================


================================================================

             1 GESTION DES APPLICATIONS

             2 CHOIX D'UNE APPLICATION

             3 GESTION DES TABLES

             4 CONTROLE

             5 ANALYSE ET APPARIEMENT

             6 EDITION DES DOSSIERS

             7 FIN DE TRAVAIL
================================================================

   Frappez le numéro correspondant à votre choix:
```

RESUME

Cette communication présente, sous ses principaux aspects, un logiciel "pour l'analyse des données collectives" SANDCO, fonctionnant sur micro-ordinateur PC ou compatible, dont l'objectif est de résoudre les différents problèmes informatiques relatifs au traitement de données sur les unités collectives.

En effet, le manque de données directement utilisables sur les unités collectives (ménage, famille, noyau familial...) est l'un des principaux obstacles au développement des recherches en "démographie de la famille"; obstacle d'autant plus regrettable que l'information de base (les relations de parenté des individus à l'intérieur de la structure collective) existe le plus souvent dans les fichiers informatiques. Mais en l'état des outils disponibles, l'utilisation de ces informations pour la création de fichier sur les unités collectives nécessite un programme informatique spécifique ou le passage par certains ordres des logiciels existants, souvent très contraignant.

Le module principal du logiciel SANDCO permet, pour chaque unité collective, de créer un enregistrement à partir des enregistrements des individus de chaque unité. L'enregistrement créé peut comprendre des variables relatives à chaque individu ou à des individus déterminés (le chef de l'unité par exemple) et des variables résumant la nature de la structure collective (type, taille, structure des différents composants, nombre d'individus pour chaque relation de parenté).

Outre le corps central de ce logiciel, sont présentés les modules complémentaires qu'ils soient destinés le plus généralement à être utilisés en amont (recodification, contrôle des données des enregistrements-individu et de la structure de leur fichier) ou en aval (appariement des fichiers, édition de dossier).

4

CONFERENCE « FEMME, FAMILLE ET POPULATION »
OUAGADOUGOU, BURKINA FASO 24-29 AVRIL 1991

CONFERENCE ON "WOMEN, FAMILY AND POPULATION"
OUGADOUGOU, BURKINA FASO APRIL 24-29, 1991

FDO Eric

Volume 1

Communications sollicitées

Commissioned papers

UEPA

Union pour l'Etude de la Population Africaine

UAPS

Union for African Population Studies